

Definition of the Different Classes of Evidence (CoE)

Articles on treatment

Class	Bias risk	Studies of Therapy	
		Study design	Criteria
I	Low risk Study adheres to commonly held tenets of high quality design, execution and avoidance of bias	Good quality RCT	<ul style="list-style-type: none"> Random sequence generation Allocation concealment Intent-to-treat analysis Blind or independent assessment for important outcomes Co-interventions applied equally F/U rate of 80%+ Adequate sample size
II	Moderately low risk Study has potential for some bias; study does not meet all criteria for class I, but deficiencies not likely to invalidate results or introduce significant bias	Moderate or poor quality RCT Good quality cohort	<ul style="list-style-type: none"> Violation of one of the criteria for good quality RCT Blind or independent assessment in a prospective study, or use of reliable data^a in a retrospective study Co-interventions applied equally F/U rate of 80%+ Adequate sample size Controlling for possible confounding^b
III	Moderately High risk Study has significant flaws in design and/or execution that increase potential for bias that may invalidate study results	Moderate or poor quality cohort Case-control	<ul style="list-style-type: none"> Violation of any of the criteria for good quality cohort Any case-control design
IV	High risk Study has significant potential for bias; lack of comparison group precludes direct assessment of important outcomes	Case series	<ul style="list-style-type: none"> Any case series design

^aOutcome assessment is independent of healthcare personnel judgment. Reliable data are data such as mortality or re-operation.

^bAuthors must provide a description of robust baseline characteristics, and control for those that are unequally distributed between treatment groups.

Determination of Overall Strength of Evidence (SoE)

After individual article evaluation, the overall body of evidence with respect to each outcome is determined based on precepts outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group and recommendations made by the Agency for Healthcare Research and Quality (AHRQ). Qualitative analysis is performed considering the AHRQ required and additional domains. The table below provides an outline of the method used to determine the final SoE.

Strength of Evidence for Existing Systematic Reviews

Level of evidence ratings for Cochrane reviews and other systematic reviews are assigned a baseline score of HIGH if RCTs were used, LOW if observational studies were used. The rating can be upgraded or downgraded based on adherence to the core criteria for methods, qualitative, and quantitative analyses for systematic reviews (there is a reference/evaluation table for this).

The following four possible levels and their definition are reported:

- High:** High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect.
- Moderate:** Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate.
- Low:** Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and likely to change the estimate.
- Insufficient:** Evidence either is unavailable or does not permit a conclusion.

All AHRQ "required" and "additional" domains^a are assessed. Only those that influence the baseline grade are listed in table.

Baseline strength: Risk of bias (including control of confounding) is accounted for in the individual article evaluations. HIGH = majority of articles Level I/II. LOW = majority of articles Level III/IV.

DOWNGRADE: Inconsistency^b of results (1 or 2); Indirectness of evidence (1 or 2); Imprecision of effect estimates (1 or 2); Sub-group analyses not stated apriori and no test for interaction (2)

UPGRADE: Large magnitude of effect (1 or 2); Dose response gradient (1)

Outcome	Strength of evidence	Conclusions and comments	Baseline	DOWNGRADE	UPGRADE
Outcome	HIGH	Summary of findings	HIGH Level I/II studies	NO consistent, direct, and precise estimates	NO
Outcome	MODERATE	Summary of findings	LOW Level III studies	NO consistent, direct, and precise estimates	YES Large effect
Outcome	LOW	Summary of findings	HIGH Level I/II studies	YES (2) Inconsistent Indirect	NO

^aRequired domains: risk of bias, consistency, directness, precision. Plausible confounding that would decrease observed effect is accounted for in our baseline risk of bias assessment through individual article evaluation. Additional domains: dose-response, strength of association, publication bias.

^bSingle study = "consistency unknown".

Articles on prognosis or risk

Class	Risk of bias	Studies of prognosis	
		Study design	Criteria
I	Low risk Study adheres to commonly held tenets of high quality design, execution and avoidance of bias	Good quality cohort ^a	<ul style="list-style-type: none"> Prospective design Patients at similar point in the course of their disease or treatment F/U rate of $\geq 80\%$^b Patients followed long enough for outcomes to occur Accounting for other prognostic factors^c
II	Moderately low risk Study has potential for some bias; does not meet all criteria for class I but deficiencies not likely to invalidate results or introduce significant bias	Moderate quality cohort	<ul style="list-style-type: none"> Prospective design, with violation of one of the other criteria for good quality cohort study Retrospective design, meeting all the rest of the criteria in class I
III	Moderately high risk Study has flaws in design and/or execution that increase potential for bias that may invalidate study results	Poor quality cohort Good quality case-control or cross-sectional study	<ul style="list-style-type: none"> Prospective design with violation of 2 or more criteria for good quality cohort, or Retrospective design with violation of 1 or more criteria for good quality cohort A good case-control study^d A good cross-sectional study^e
IV	High risk Study has significant potential for bias; does not include design features geared toward minimizing bias and/or does not have a comparison group	Poor quality case-control or cross-sectional Case series ^d	<ul style="list-style-type: none"> Other than a good case-control study Other than a good cross-sectional study Any case series^f design

^aCohort studies follow individuals with the exposure of interest over time and monitor for occurrence of the outcome of interest.

^bApplies to cohort studies only.

^cAuthors must consider other factors that might influence patient outcomes and should control for them if appropriate.

^dA good case-control study must have the all of the following: all incident cases from the defined population over a specified time period, controls that represent the population from which the cases come, exposure that precedes an outcome of interest, and accounting for other prognostic factors.

^eA good cross-sectional study must have all of the following: a representative sample of the population of interest, an exposure that precedes an outcome of interest (e.g., sex, genetic factor), an accounting for other prognostic factors, and for surveys, at least a 80% return rate.

^fA case-series design for prognosis is one where all the patients in the study have the exposure of interest. Since all the patients have the exposure, risks of an outcome can be calculated only for those with the exposure, but cannot be compared with those who do not have the exposure. For example, a case-series evaluating the effect of smoking on spine fusion that only recruits patients who smoke can simply provide the risk of patients who smoke that result in pseudarthrosis but cannot compare this risk to those that do not smoke.

Definitions of the Different Levels of Evidence for Reliability Studies

Level	Study type	Criteria
1	Good quality study	<ul style="list-style-type: none"> Broad spectrum of persons with the expected condition Adequate description of methods for replication Blinded performance of tests, measurements or interpretation Second test/interpretation performed independently of the first
2	Moderate quality	<ul style="list-style-type: none"> Violation of any one of the criteria for a good quality study
3	Poor quality study	<ul style="list-style-type: none"> Violation of any two of the criteria
4	Very poor quality study	<ul style="list-style-type: none"> Violation of all three of the criteria